NAVAL HEALTH RESEARCH CENTER

FIELD APPLICATIONS OF COGNITIVE ASSESSMENT BATTERIES:

INITIAL TESTS OF ALTERNATIVE MEASUREMENT MODELS

R. R. Vickers, Jr.

J. F. Kusulas

Report No. 92-8

NAVAL HEALTH RESEARCH CENTER
P. O. BOX 85122
SAN DIEGO, CALIFORNIA 92186 – 5122

NAVAL MEDICAL RESEARCH AND DEVELOPMENT COMMAND
BETHESDA, MARYLAND

Field Applications of Cognitive Assessment Batteries: Initial Tests of Alternative Measurement Models

Ross R. Vickers, Jr.
Jeffrey W. Kusulas

Cognitive Performance and Psychophysiology Department
Naval Health Research Center
P.O. Box 85122
San Diego, CA 92186-5122

Technical Report 92-8 supported by the Naval Medical Research and Development Command, Navy Medical Command, Department of the Navy, under Research Work Unit MM33B30.001-6104. The views presented are those of the authors and do not reflect the official policy of the Department of the Navy, Department of Defense, nor the U.S. Government.

Summary

Military personnel often perform cognitively demanding tasks under challenging conditions. Cognitive assessment batteries have been developed to assess the sensitivity of cognitive functions to conditions such as heat, cold, and heavy physical work. Recent advances in computer technology make it possible to employ these batteries in field studies of operational exposures to stress. However, field studies typically must be performed with the constraint that interference with the operational mission be minimized. Satisfying this constraint often means that typical laboratory controls are difficult or impossible to achieve in the field. This lack of control can pose problems when interpreting the results obtained with the tests, so it is important to determine how field testing conditions affect the results obtained with cognitive batteries.

The present study evaluated the effect of being unable to follow recommendations that a series of practice sessions be conducted with cognitive tests before collecting experimental data. This recommendation is based on previous studies showing that individual differences in performance are unstable for early practice sessions. To date, the bases for the evident instability have not been investigated. Further information on this topic could help define alternative research designs to ensure valid conclusions from field studies.

Structural equation modeling procedures were applied to data reported by Kennedy, Dunlap, Jones, Lane, and Wilkes (1985) for computerized versions of the Sternberg Memory Test, Simple Reaction Time, the Manikin Test, Rate of Tapping (with separate tests for preferred hand, non-preferred hand, and both hands together), the Code Substitution test, the Grammatical Reasoning test, and Pattern Comparison test. Major findings were:

(a) All measures except the Sternberg Memory Test and Grammatical Reasoning had constant true score variance. The Sternberg Memory test had constant true score variance except for session 2. Grammatical Reasoning required a two-dimensional model to fit the data.

(b) Error variance was constant across sessions for Code Substitution, Pattern Recognition, Preferred Hand Tapping, Non-preferred Hand Tapping, and Sternberg Memory. Other tests had variable error components, primarily due to greater error for early sessions.

(c) Residual covariances between scores for temporally adjacent sessions generally were estimated to be zero.

(d) The basic models for Code Substitution and the Manikin test replicated in independent samples.

The presence of constant true score variance implies that the tests considered generally were valid from the first session onward and can be used in one-test and two-test research designs. However, research designs also must ensure that sample sizes are large enough to compensate for the tendency for the measurement precision of these tests to be lower in early sessions than in later sessions. The exception to these conclusions was the Grammatical Reasoning test which does not appear well-suited to brief studies.

Introduction

Background

Maintaining cognitive functioning under demanding conditions is necessary to ensure effective performance and safety in many work settings. Standardized tests of cognitive functioning have reliable, generalizable relationships to performance for at least some types of jobs (McHenry, Hough, Toquam, Hanson, & Ashworth, 1990; Schmitt, Gooding, Noe, & Kirsch, 1984). Adverse environmental conditions can change the level of cognitive functioning (Hockey, 1986), but the significance of these effects for real work situations is uncertain because much of the available evidence derives from laboratory studies of simulated stresses. Portable computerized cognitive testing batteries now make it possible to standardize cognitive testing in field settings where it would have been difficult, if not impossible, to conduct similarly controlled studies prior to the advent of this technology (Kennedy, Dunlap, Jones, Lane, & Wilkes, 1985).

Research Issues

Field studies differ from laboratory studies in many ways, which are largely because of less control over research conditions by the researchers. Efficient transfer of cognitive assessment technologies from the laboratory to field studies presents some problems associated with this loss of control. The study reported in this paper is part of a program to evaluate the effects of research design modifications which may be required to adapt to field conditions. The objective of these evaluations is to provide empirically-based guidelines for research design trade-offs in field studies.

Limited periods of subject availability represent one important constraint encountered in field studies. In these settings, subject availability often is constrained by requirements imposed by the subject populations' regular work or training. As a result, the number of data collection sessions must be kept to a minimum to avoid interfering with operatonal requirements.

The practical requirement of keeping data collection to a minimum conflicts with the ideal research design which would include a series of familiarization sessions to ensure appropriate psychometric characteristics for the tests (Bittner, Carter, Kennedy, Harbeson, & Krause, 1986). These recommendations are based on well-established psychometric considerations (Jones, 1979; Kennedy, Carter, & Bittner, 1980), but there is a recognized need to determine whether scores based on as few as one or two testing sessions are valid as a means of extending the applicability of computerized cognitive assessment batteries (Thorne, Genser, Sing, & Hegge, 1985). Some recent treatments of cognitive tests apparently assume the tests are valid from the first administration onward (AGARD Working Group 12, 1989), but it is desirable to test this assumption if possible.

Conceptual Approach

Tests as Mixtures. The conceptual approach to cognitive performance measurement in this study assumes that test scores represent the combined influences of three general sources of variance (Bentler & Woodward, 1980). The first source of variance is the cognitive ability(ies) the test is designed to measure. The second source of variance is transient factors affecting a single session (e.g., some distraction occurring during testing). The third source of variance is systematic influences on performance other than the cognitive ability the test is designed to measure. For example, Ackerman (1987, 1988) has shown that general intelligence and psychomotor abilities affect some cognitive performance measures. James, Mulaik, and Brett (1982) refer to these three potential sources of performance differences as "true score variance," "error variance," and "disturbance variance." In the following discussion, these terms will be employed to label sources of variance, except that "contamination variance" will be substituted for "disturbance variance." This substitution is made in the belief that the latter term more graphically conveys the present concern about possible invalid conclusions which might result from using contaminated measures.

The present study was undertaken because the statistical basis for existing recommendations regarding use of cognitive assessment batteries does not specify the bases for the instability of individual differences during early sessions in a series. Jones (1979) developed a statistical test which can be applied to scores from a series of test sessions to identify the session after which individual differences in performance are stable. This test has been extensively used in developing recommendations about the use of cognitive tests (Bittner et al., 1986).

Jones' (1979) test was not designed to examine the reasons for instability prior to stabilization. This fact is important because all three sources of variance in test scores noted above can contribute to the instability characteristics of early sessions. However, these sources of variance do not necessarily have to become sources of invalid research conclusions. For reasons considered below, recommendations based on Jones' (1979) test may be conservative.

A case for valid measurement of ability prior to stabilization of individual differences can be made by examining the implications of changes in each of the sources of variance underlying test scores. If experience with the test helps develop the ability being measured, differential growth curves for individuals will produce imperfect correlations between trial-to-trial individual differences. However, since the actual source of variance would still be the ability(ies) the test was designed to measure, these changes will not produce invalid variance. Instead, the unstable correlations between individual differences will be a factor that will affect the legitimacy of statistical conclusions if the data are analyzed by a simple repeated measures analysis procedure. Proper application of multivariate analysis procedures can provide an appropriate statistical basis for research conclusions (in terms of average changes in performance) even when individual differences in growth or learning curves occur. Accurate estimates of stressor effects should be provided if these analysis procedures are combined with research designs which include appropriate controls for learning curve effects (e.g., control groups or counterbalanced research designs).

Changes in error variance represent a second possible basis for the instability of early session correlations. These changes will alter the measurement precision or reliability of the test (Lord & Novick, 1968), but error variance is random by definition. As a result, error variance does not affect the estimated magnitude of the effects for experimental variables. Applying this logic to field studies of cognitive ability(ies), appropriate conclusions can be reached by performing statistical power analyses (Cohen, 1968), then adjusting either the sample size or the significance criterion to allow for the expected magnitude of measurement error.

Contamination variance means that changes in factors other than the cognitive ability(ies) of interest can produce changes in overall test performance across sessions. This source of variance is troublesome because change resulting from contaminants might be mistakenly interpreted as change due to some experimental factor of interest. For example, if motivation were a significant component of performance on a given task, a decrement or increment in motivation that coincided with the onset of experimental conditions would lead to a corresponding change in performance without a concomitant change in the underlying cognitive ability. A researcher might conclude, therefore, that the experimental condition impaired or enhanced cognitive functioning when it really affected motivation. If one were concerned solely with predicting performance, this interpretive error would be unimportant. However, the error could be critical for programs to minimize the effects of the experimental condition. An ability interpretation might lead to the decision to implement further training which could actually exacerbate the problem, whereas a motivational interpretation could lead to programs designed to improve this component of the performance equation.

It is important to note that even contamination does not preclude valid research conclusions. If contaminants can be identified, research designs can include measures of the contaminating factors. Valid estimates of effects then can be obtained by employing statistical controls for the effects of these extraneous variables on test performance. While this approach may not appear ideal relative to a pure experimental design, it may be the only option available in many instances.

The arguments presented above lead to the conclusion that instability in test scores does not preclude valid conclusions from studies that employ only one or two testing sessions. The effect of each source of variance can be controlled by appropriate design and analysis procedures, so studies conducted without familiarization sessions are feasible in principle. It should be equally clear that there is an increased risk of invalid conclusions if appropriate designs are not implemented to transform principle to reality. The development of appropriate research designs depends on having adequate understanding of the reasons for the initial instability in the test scores because different research design and analysis strategies must be implemented to deal with each type of problem.

Study Objectives and Approach

The preceding arguments illustrate that with appropriate information about the bases for the instability of individual differences in early test sessions it is possible to develop research designs using cognitive tests that yield valid conclusions even when data collection must be limited to one or two testing sessions. However, the magnitude of the contribution of each source of variance must be known to define the appropriate research design and statistical analyses required to ensure that such studies produce valid conclusions. The present study applied the conceptual model described above to evaluate a set of cognitive tests taken from the

Uniform Tri-Services Cognitive Performance Assessment Battery (UTC-PAB; Englund et al., 1987).

Appropriate statistical procedures are needed to estimate the contributions of true score, contamination, and error variance. Recent developments in structural equation modeling (SEM; Bentler, 1989; Joreskog & Sorbom, 1981) can be adapted to this purpose. In the present case, the application of these techniques began with a model which assumed that the cognitive tests conform to Lord and Novick's (1968) definition of parallel measures from the initial session onward. This model implies equal true score variance for all sessions and equal error variance for all sessions. A parallel tests model also implies that the source(s) of true score covariation are the same for all sessions, so scores from different sessions will be uncorrelated after taking these constant factors into account.

The parallel measures model was not expected to fit the data given past evidence that several sessions are needed to stabilize test scores (Kennedy et al., 1981). The basic purpose of the parallel tests model, therefore, was to serve as a frame of reference for testing hypotheses about the sources of variance underlying instability in the early sessions.

The general analysis approach was to fit a structural equation model representing the parallel tests model, then remove specific constraints implied by that model to determine the basis for misfit between the model and the data (see Analysis Procedures). If factor loadings had to be modified, the assumption of equal true score variance was questionable. If error terms had to be modified, the assumption of equal error variance was questionable. If residual covariance constraints had to be modified, the assumption that contamination was near zero was questionable. The overall implications of the study for research designs, therefore, depended on the pattern of modifications required to bring the basic measurement model in line with the data.

Method

Data Source

Kennedy et al. (1985) presented tables of the means, standard deviations, and intercorrelations of cognitive and psychomotor tests for a sample of 25 college students (8 males, 17 females) between 18 and 25 years of age.

Performance Tasks.

The data analyzed in this report were limited to results obtained by Kennedy et al. (1985) for tasks subsequently incorporated into the Uniform Tri-Services Cognitive Performance Assessment Battery (UTC PAB) and administered by computer in their study. These tasks included:

(a) Sternberg Memory Test (Sternberg, 1966): Subjects were presented with a series of probe letters after commiting four target letters to memory. Subjects indicated whether the probe letter was or was not one of the target letters presented prior to a series of probe sessions. Scores were expressed as number right minus number wrong divided by time on task.

(b) Reaction Time: An auditory signal preceded a visual signal by three seconds. The subjects pressed a key as quickly as possible following the onset of the visual signal.

(c) Manikin Test (Benson & Gedye, 1983): A simulated human figure is shown from a full front perspective or a full back perspective. The figure is holding two easily distinguished objects, one in each hand. One of the two objects is shown below the figure, and the subject is required to indicate whether it is the left or right hand of the figure that is holding this object. Score was the number of correct responses minus the number of errors divided by time on task.

(d) Tapping: Subjects were required to alternate presses on a pair of specified keys on the computer keyboard. This test was done using the preferred hand, the nonpreferred hand, and both hands at once. Score was the number of alternating key presses made in the allotted time.

(e) Code Substitution (Ekstrom, French, Harmon, & Derman, 1976): Digits are randomly assigned to nine letters. A test letter is then presented and the subject is required to respond with the corresponding digit. Scoring was the number correct minus the number wrong.

(f) Grammatical Reasoning (Baddeley, 1968): Statements are provided about the sequence of two letters, A and B. Each statement pertains to which letter comes first or last in a sequence that is shown on the screen at the same time as the statement. Five grammatical transformations are provided regarding the relationship between two letters, A and B. Subjects were required to indicate whether the given statement was true or false relative to the pattern of letters actually shown. Score was the number attempted minus the number wrong. Although not specifically stated, the test apparently was timed.

(g) Pattern Comparison (Klein & Armitage, 1979): Two patterns of dots were presented simultaneously and subjects were required to indicate whether the patterns were the same or different. Scores were number of patterns attempted minus number wrong.

Analysis Procedures

Construction of Alternative Models. Structural equation models (SEMs) were estimated with LISREL VI (Joreskog & Sorbom, 1981). Three models were estimated for each performance measure, as described below:

(a) Null Model. This model assumed that individual differences in scores from each testing session were solely the product of random factors affecting that session. The random factors were assumed to be independent across trials and to cumulate to the same total variance in scores for each trial. This model was operationalized by stipulating that there was a single latent trait represented in the data with a zero loading ("lambda" in LISREL terminology) on that trait for the scores from each session. In addition, the residual variances ("epsilons" in LISREL terminology) were assumed to be equal across all sessions. The model, therefore, was estimated with the following constraints:

lambda = 0 for each session

$\text{epsilon}_i = \text{epsilon}_j$ for any sessions $\underline{i}$ and $\underline{j}$

This model will fit the data perfectly if the covariance matrix being analyzed consists of a constant variance on the diagonal and zeros in the off-diagonal positions.

(b) Parallel Tests Model. The primary alternative to the null model was a parallel tests model which was a more complex model based on several assumptions. One assumption was that scores for each trial of each cognitive/psychomotor test were influenced by one or more stable attributes of the individual. These attributes were the source of true score variance which was constant across trials. The presence of non-zero true score variance was represented in the model by a non-zero factor loading (i.e., lambda) for each session. The fact that the loadings were assumed to be constant was operationalized by imposing the constraint that lambda was identical for all sessions when the computations were performed to estimate the model parameters. A second assumption was that the stable attributes which were the basis for true score variance were the only source of covariation between scores from different sessions. This assumption was equivalent to asserting that other sources of variance were uncorrelated from one session to the next. The last assumption in the model was that the variance associated with random influences was constant across all sessions just as was assumed in the null model. In the same terms that were used to describe the null model, the parallel tests model was estimated with the following constraints:

lambda = k for each session, where _k_ is not equal to 0

$\text{epsilon}_i = \text{epsilon}_j$ for any sessions _i_ and _j_

These assumptions constituted an SEM which conformed to Lord and Novick's (1968) definition of parallel measures as measures which were comprised of equal true score variance and equal error variance. The model would fit the data perfectly if the covariance matrix representing the data took the form of a single constant value in the off-diagonal positions and a single constant value in the diagonal positions which was greater than or equal to the value in the off-diagonal positions.

Scaling the Models. SEMs require that some element of the model be fixed to establish the unit of scaling for each latent trait. The alternatives are to fix the scaling for a given indicator (e.g., an item in a scale or, in this case, one of the sessions) and scale other indicators in terms of their relative contributions to the latent trait variance or to fix the variance of the latent trait (Joreskog & Sorbom, 1981). The former approach produces models in which the factor loading for one indicator is not estimated, and, in the present application, would require that all other factor loadings be fixed at whatever value was stipulated for the chosen indicator. This equivalence would be necessary to be consistent with the constraint of equal loadings in the parallel tests model. As a result, the approach of fixing the loading for a single indicator was not informative in the present SEM application. Fixing the variance of the latent trait to establish the measurement model permits factor loadings to be estimated for each indicator and, in the present case, provides the required estimates of true score variance for each session in the parallel tests model.

Model Fit Indices. The stepwise process stopped when the removal of an additional constraint produced a model with a smaller parsimony-adjusted Tucker and Lewis (1973) fit index value than the preceding model.[1] The Tucker-Lewis index (hereafter, TLI) represents the proportion of the greater than chance covariation between measures that is accounted for by the model. The formula for the TLI is

$$\text{TLI} = (\text{Chi-square}_N/\text{df}_N - \text{Chi-square}_T/\text{df}_T)/(\text{Chi-square}_N/\text{df}_N - 1)$$

where the chi-squares are values computed by the program based on the residual covariance between measures after the covariance accounted for by the model is subtracted, df indicates degrees of freedom, N indicates the null model, and T indicates the target model, the particular alternative to the null model that is being evaluated at that time.

Parsimony adjustments were applied to the TLI based on philosophical (Mulaik et al., 1989) and statistical (Bentler & Mooijaart, 1989) considerations. In general, SEMs with more parameters to be estimated will fit a given data set better than simpler models, so reliance on simple measures of goodness-of-fit will tend to lead to the retention of more complex models at the cost of parsimony. Applying an adjustment to take into account differences in the complexity of alternative models provides a guard against this problem, as discussed by Mulaik et al. (1989), and provides parameter estimates which have smaller sampling variance than those derived in more complex models (Bentler & Mooijaart, 1989). The resulting index, referred to below as the adjusted TLI (ATLI), is computed by multiplying the TLI by df_T/df_N.

The choice of a null model obviously affects the size of the goodness-of-fit indices, so the construction of alternative null models would have produced different goodness-of-fit results. The chosen null model seemed the most appropriate for the present purposes for two reasons. First, this model was the most constrained model that could be fitted plausibly to the data. Second, the parameters of this model formed a nested subset of the parameters estimated in the substantive models considered, thereby fulfilling a requirement for the use of nested hierarchical tests of fit between alternative models (see Bentler & Bonett, 1980, for a discussion of nested models and the evaluation of fit).

Model Modification. The parallel measures model was not expected to fit the data perfectly, so procedures for examining the misfit between the parallel measures model and the data were needed. The approach adopted was to examine modification indices for individual parameters (i.e., the factor loading for scores on a particular session, the error variance for a particular session or the residual covariation between two particular sessions)[2]. Modification indices are estimates of how much the chi-square statistic for the model would improve if a single constrained parameter were estimated freely. If the modification indices for all the constrained parameters were small, the parallel tests model would be accepted.[3]

If any modification index for the factor loadings (lambdas), errors (epsilons) or correlations between errors for adjacent sessions exceeded a chi-square criterion discussed below, the constrained parameter with the largest modification index value was freed. Values for the free parameters in the revised model were estimated, the improvement in fit was determined, and the resulting model was inspected to see whether additional constraints should be removed using the fit indices described below.

The model modification approach adopted here is comparable to automatic modification procedures in structural equation modeling programs (Bentler, 1989; Joreskog & Sorbom, 1981). The major difference is that the present approach focused on a subset of the possible parameters that could be modified. This point is noteworthy because the reliability and accuracy of automatic modification procedures in uncovering the correct model underlying a data set is a point of some dispute (Bentler & Chou, 1990; Joreskog & Sorbom, 1990; Spirtes, Scheines, & Glymour, 1990a,b). The present approach attempted to increase the accuracy of the search for a final model by constraining the procedure to look for the most likely and readily interpretable parameter changes. However, the modifications were intended to explore areas of misfit between the data and the parallel tests model. The results are not intended as definitive models, but as plausible alternatives to be replicated in future studies.

Results

Fit of the Parallel Tests Model

The null model produced ATLI values in excess of .75 for 6 of 9 measures (Table 1). Two-hand Tapping fell slightly short of this criterion and both Logical Reasoning and Four-Choice Reaction Time were poorly approximated by the parallel tests model.

Trends in Misfit for the Parallel Tests Model

A summary picture of the effects of specific constraints on misfit between the parallel tests model and the data is provided in Table 2. The table entries are the summed modification indices for each constrained value under the parallel tests model. The values given are the result of aggregating the individual chi-squares for that constraint across the nine tests. Since each entry is the sum of 9 presumably independent chi-squares, tests for significance based on 9 degrees of freedom have been provided. Given that the modification indices are estimates of the minimum chi-square change that would be obtained by freeing the constrained parameter, the significance values provided in the table probably are conservative. However, the uncertainty about the appropriate interpretation of these cumulative chi-squares as guides to the statistical significance of modifying parameter constraints is not critical because Table 2 is intended more as a descriptive summary of the loci for misfit between model and data than as a basis for statistical inference.

Summing the columns of Table 2 provides an index of the effect of a given set of constraints (i.e., constraints on lambdas, epsilons, and adjacent session covariances) on the fit between the data and the parallel tests model. Using these sums as a guide, the equal error constraint was the major source of misfit (sum = 54.75), followed by the uncorrelated error constraint (sum = 25.22), with the true score constraint a relatively minor source of misfit (sum = 15.17).

The same picture of misfit between the model and the data can be illustrated in a different fashion by considering the modification indices for particular tests. Excluding Grammatical Reasoning on the grounds that a single-factor model was not reasonable for this test (see pp. 17-19), only 4 of 80 modification indices for the equal true score constraint exceeded 3.84, compared to 15 of 80 for the equal error constraint, and 14 of 72 for the zero residual covariances between adjacent sessions constraint.

Table 1
Summary of Essex Data Results

	Chi-Square for:			
Task	Null	Parallel	TLI	ATLI
Manikin	427.29	134.36	.78	.76
Sternberg Memory	296.29	83.75	.87	.85
Pattern Recognition	432.55	103.48	.86	.85
Code Substitution	285.27	70.07	.92	.91
Grammatical Reasoning	369.11	206.01	.51	.50
Four-Choice Reaction	253.28	193.69	.28	.28
Tapping Tests:				
Preferred Hand	493.33	97.45	.90	.88
Non-Preferred Hand	562.55	123.17	.86	.84
Two-Hand	460.98	176.02	.69	.68

NOTE: The Null Model had 54 degrees of freedom and the Parallel model had 53 degrees of freedom. "TLI" refers to the raw Tucker-Lewis index. "ATLI" refers to the parsimony-adjusted Tucker-Lewis Index.

Table 2
Summary of Cumulative Modification Indices for Different Sessions

	Cumulative Chi-Square for Constraint of:					
Test	Equal True Score		Equal Error		Uncorrelated Error	
Session	Average	Maximum	Average	Maximum	Average	Maximum
1	3.45**	13.78	29.57**	121.24		
2	2.10*	9.58	6.39**	45.00	8.06**	26.60
3	1.51	5.49	2.91**	9.80	1.10	4.40
4	1.79	10.39	4.72**	24.30	2.77*	14.97
5	.98	3.29	2.54*	5.06	1.72	4.78
6	.45	1.21	1.30	4.34	1.19	4.50
7	1.02	3.11	2.04*	7.84	.85	2.13
8	2.58*	14.63	1.50	4.01	2.89*	13.88
9	.75	3.68	1.31	6.94	3.02*	12.27
10	.54	2.29	2.47*	8.45	3.62*	9.82

* p < .05 (Critical value = 1.88, 9 df)
** p < .0056 (Critical value = 2.60, 9 df with Bonferroni adjustment)

The probability of getting the observed number of modification indices which exceed the 3.84 criterion can be determined by noting that a modification index can be regarded as a chi-square with 1 degree of freedom. The sampling distribution for chi-squares therefore can be used to determine that the proportion of modification indices expected to be greater than or equal to 3.84 is .05. If this proportion is taken as the probability of "success" when comparing observed chi-squares to the criterion, the binomial probability of obtaining the observed number of results which exceed the criterion can be computed for each set of modification indices. Under this assumption, the probability of obtaining 4 or more modifications indices which exceed 3.84 in a set of 80 such indices is p > .57. By contrast, the probability of obtaining 15 indices in excess of 3.84 in a set of 80 is p < .001. The probability of 14 indices in excess of 3.84 in a set of 72 also is p < .001. By this test, the frequency of significant modification indices associated with the equal lambdas constraint was essentially chance, but the frequencies of significant modification indices associated with the equal error constraint and the zero covariance constraint both were much greater than chance.

In general, the cumulative chi-squares presented in Table 2 tended to be inflated by a few large values. Fifteen of the 29 constraints produced cumulative modification indices which were greater than the critical value for p < .05. However, only 4 of these 15 cumulative chi-squares would have remained significant (p < .05, critical value = 1.94, 8 df) if the task with the maximum chi-square for that constraint had been excluded from the calculations. The four constraints that would have produced significant chi-squares even with the highest modification index deleted were for constraints involving session 1 or session 2 of the data collection. In particular, the cumulative chi-squares which would have been significant even with the largest single value excluded were the equal lambda constraint for session 1, the equal error constraints for sessions 1 and 2, and the residual covariance constraint for session 1 with session 2.

The misfit between the data and the parallel tests model was not randomly distributed across the tests. Grammatical Reasoning produced 10 parameter constraints with chi-square values in excess of 3.84, as might be expected given the poor overall fit of the parallel tests model for this test (Table 1). However, misfit was not distributed randomly even when attention was restricted to the other eight tests. Reaction Time accounted for 7 of 15 large modification indices for the equal error constraint, and the three tapping tests accounted for 9 of 14 large modification indices for the uncorrelated errors constraint.

Modification of the Basic Model for Individual Cognitive Tests

Modifications to the basic parallel tests model were made on a test-by-test basis following the stepwise process described in the Analysis Procedures (pp. 10-13). The results of these modifications are summarized in Table 3. The important general trends were:

(a) The parallel test model was retained for Code Substitution, Pattern Recognition, and Non-preferred Tapping.

(b) Except for Grammatical Reasoning, which is discussed further below, at most 2 constraints were modified for any test.

(c) Five of the seven modifications were for equal error constraints.

(d) The error correlation for sessions 8 and 9 for Preferred Hand Tapping and the unequal error for session 4 Reaction Time would not pose problems for research designs involving one or two test sessions.

Grammatical Reasoning. Sequential modification of the parallel tests model for Grammatical Reasoning did not identify a viable model for the data. Freeing a constrained parameter to eliminate a large modification index improved the fit of the model to the data for the session associated with that index in the preceding model, but this improvement was accompanied by increases in the size of other modification indices. After an extensive series of modifications, the analyses had not converged to define a reasonable model for Grammatical Reasoning. This problem made it reasonable to consider the possibility that the basic unidimensional model was inappropriate for this task.

Examination of the pattern of correlations between sessions, the modification indices for the parallel tests model, and the residual covariances generated by fitting the parallel tests model suggested that the reason for this problem was that an alternative model which treated early and late sessions as indicators of partially independent constructs was worth considering. Based on this observation, a model which assumed that early sessions (1-4) defined a performance factor that was partially independent of performance differences in later sessions (5-10) was fitted to the data. Factor loadings (lambdas) were assumed to be constant within dimensions and error variance (epsilon) was assumed to be constant across all sessions. The two dimensions were assumed to be correlated.

Table 3
Summary of Model Modifications

	Chi-Square for:				
		Improvement			Chi-Square
	Model	in Fit	TLI	ATLI	%
Manikin					
(a) Parallel	134.36	292.93	.78	.76	88.7
(b) Error 1	97.18	37.08	.87	.84	100.0
Sternberg Memory					
(a) Parallel	83.75	212.54	.87	.85	93.4
(b) Lambda 2	73.88	9.87	.91	.87	97.7
Code Substitution					
(a) Parallel	70.07	215.20	.92	.91	100.0
Pattern Recognition					
(a) Parallel	103.48	329.07	.86	.85	91.7
Reaction Time					
(a) Parallel	193.69	59.59	.28	.28	34.9
(b) Error 1	137.33	56.36	.56	.53	67.8
(c) Error 4	96.52	40.81	.76	.72	91.7
Grammatical Reasoning	See Text				
Tapping Tests:					
Preferred Hand					
(a) Parallel	97.45	395.88	.90	.88	92.1
(b) Error Corr 8-9	84.56	12.89	.92	.89	95.1
Non-Preferred Hand					
(a) Parallel	123.17	439.38	.86	.84	93.6
Two-Hand					
(a) Parallel	176.02	284.96	.69	.68	76.2
(b) Error 1	158.90	17.12	.73	.70	80.8
(c) Error 2	107.44	51.46	.85	.81	94.6

NOTE: "Parallel" = parallel tests model. "Lambda," "Error," and "Error Corr" = type of constraint freed. Associated numbers indicate the session(s) involved, e.g., "Error 1" refers to a model produced from the preceding model by removing the equal error constraint for the first session. Improvement of fit is the change in chi-square from the prior model with the null model from Table 1 as the prior model for the Parallel model. "TLI" is the Tucker-Lewis index; "ATLI" refers to the parsimony-adjusted TLI. The "Chi-Square %" = (Null Model Chi-Square - Current Model Chi-Square)/ (Null Model Chi-Square - Minimum Chi-Square) where the Minimum Chi-Square is the chi-square obtained by continuing the model modification process until all modification indices were less than 3.84.

The 2-factor model substantially improved the goodness-of-fit relative to the one-dimensional model (Table 4), but still did not accurately represent the pattern of associations between scores for different sessions. Examination of the modification indices suggested that the original assignments of session 2 and session 8 scores to the early and late factors, respectively, had been inappropriate choices. These indices suggested that the initial assignments should be reversed for these two sessions, so two models were fitted to make these changes. The changes substantially improved on this initial model, but even this model provided only a modest absolute level of fit between the model and the data (ATLI = .69).

Table 4
Structural Model for Grammatical Reasoning

	Chi-Square for:				
		Improvement			Chi-Square
	Model	in Fit	TLI	ATLI	%
Parallel One-Factor	206.11	163.10	.51	.50	54.9
Parallel Early-Late	157.69	48.42	.64	.61	71.1
Session 2 Modification	139.00	18.69	.70	.67	77.4
Session 8 Modification	125.66	13.34	.74	.69	81.9

The correlation between the two dimensions in the alternative Grammatical Reasoning models is interesting as an index of the extent to which the data deviated from the unidimensionality of the original parallel tests model. The values of this correlation varied from one model to the next, ranging from .76 for the initial two-dimensional model down to .60 for the final model in Table 4. Based on these findings, a unidimensional model for performance on the Grammatical Reasoning test is unreasonable and the simplest alternative model with early and late performance components also seems questionable.

Replication across Samples

Multivariate analyses performed in small samples must be regarded with caution despite some evidence that parameter values for models can be accurately recovered with such samples (McArdle, 1991). Small sample sizes are associated with greater sampling variability which permits wider ranges of chance variation in estimated parameters. Small sample sizes also enhance the impact of any outlier data points on the analyses. The results obtained with other published data sets, therefore, were compared to the preceding findings to evaluate the replicability of the results obtained in the initial analyses. The expectation was that modifications to the basic parallel tests model would be less likely to replicate than would the model itself.

Code Substitution. An independent set of performance data for the Code Substitution test was provided by Pepper, Kennedy, Bittner, and Wiker (1980; referred to as "Pepper Data" in Table 5). These authors reported the correlation matrix for 10 sessions on this test for 18 U.S. Navy personnel. The standard deviations across sessions were plotted in a figure, but not reported in tabular form. This fact meant that standard deviations could not be determined with sufficient accuracy to permit reconstruction of the covariance matrix. Analyses to compare results in this sample to those in the Essex sample, therefore, were conducted with the correlation matrices.

Table 5
Cross-validation of Structural Models for Code Substitution

	Parallel Model:		Chi-square for:			ATLI for:	
	Lambda	Epsilon	Null	Parallel	Cross-Validation	P	C-V
Code Substitution							
Pepper Data	.812	.340	247.86	115.93	117.68	.66	.68
Essex Data	.840	.294	283.03	66.34	68.58	.92	.94
Manikin							
Carter Data	.837	.300	298.75	130.64	144.89	.66	.64
Essex Data	.892	.205	425.05	132.77	146.86	.77	.76
df =			54	53	55		

The replicability of the initial findings was demonstrated in several ways (Table 5), including the close comparability of the lambda and epsilon estimates derived from the two data sets. The chi-squares for the sample-specific parallel model and the cross-validated model cannot be treated as incremental fit indices because these two models do not form a nested hierarchical sequence as required for these tests. However, if the chi-square for the null model in each sample is used as a reference point, the cross-validated model achieved 98.7% of the reduction in the chi-square that the sample-specific model did when both were applied to the Pepper et al. (1980) data. The comparable figure for the Essex data was 99.0%.

Manikin Test. Carter and Woldstad (1985; referred to as "Carter Data" in Table 5) reported the intersession correlations for 10 sessions with the Manikin test for 20 U.S. Navy enlisted men. Based on the mean levels and standard deviations for this sample, it appears that a longer version of the test was used in this study than was used by Kennedy et al. (1985). The increased length of the test appeared to increase the variance of the scores substantially, as would be expected, so a direct replication of the analysis of covariance in this report was not possible. The cross-validation was conducted with correlation matrices to provide at least some assessment of the cross-sample reliability of the present findings.

As indicated in Table 5, the two samples produced comparable results. The findings were less consistent across samples than was the case for Code Substitution, but the cross-validated model still accounted for 92.1% of the covariance explained by the sample-specific model when applied to the Carter and Woldstad (1985) data. The comparable figure was 95.2% for the Essex data.

Discussion

The results of these analyses indicate that the cognitive and psychomotor tests generally have constant true score variance and zero residual covariances once covariance due to true scores is taken into account. The findings were mixed with regard to error variance with about half of the tests having constant error variance and the remainder having variable error variance. When error variances did differ across sessions, the differences were most likely to occur for the first testing session. Given this general pattern of results, the tests examined can be characterized as parallel measures from the first session onward if they had constant error variance and as tau-equivalent tests if they had differential error variance across sessions (Lord & Novick, 1968). The practical implication is that these tests are valid from the first session, but may provide more precise measurement of individual differences after several testing sessions than they do at the first session. Research designs involving only one or two testing sessions, therefore, must be appropriately sensitive to the expected lower measurement precision, but this problem can be dealt with by power analyses to guide the sample size determinations or by adjustment of significance levels to be sensitive to effect sizes that are of interest.

There were exceptions to the above generalizations, but they clearly applied to 7 of the 9 tests considered and an argument can be made for their extension to the Sternberg Memory test. The Sternberg Memory test would have satisfied the requirements for a parallel tests model except for evidence that the true score variance for session 2 differed from that of other sessions. In the context of the current finding that there typically are no differences in true score variance across sessions, this single violation of the equal true score constraint could be explained as a chance finding. However, an alternative explanation would be consistent with the typical pattern of increasing correlations between test scores across repeated testing sessions found in studies of cognitive and psychomotor performance (Bittner et al., 1986). The monotonicity of the typical trend suggests that the underlying mechanism(s) of change operate from the first trial onward. If such mechanisms actually exist and affect performance on the Sternberg Memory Test, the true score variance on session 2 might really differ from that on later trials, and the equivalence of the estimated session 1 true score variance with the true score variance on later trials would be the anomalous finding. In the overall context of the present study, the more parsimonious explanation would be that the deviation of session 2 true score variance from that on other trials is the chance finding. Unless further study demonstrates that this is a reliable phenomenon, 8 of 9 tests can be fitted by the parallel or tau-equivalent models.

The Grammatical Reasoning test data were not fitted well by any simple unidimensional measurement model. Based on the present data, the Grammatical Reasoning is not a good candidate for studies when extensive familiarization with the test is not possible, but this general finding may be a peculiarity of the specific sample studied. Other samples have produced correlation matrices with more stable individual differences for this test (Carter, Kennedy, & Bittner, 1981). Furthermore, even if the present results replicate, what appear to be problems with Grammatical Reasoning actually may reflect strengths of the test, at least for some purposes. This test seems inherently more complex than the others considered here. Complex tests would seem more susceptible to disruption by stress on the grounds that there are more ways that performance can go wrong and that it is more difficult to make responding automatic. At the same time, complex tests may be more valid predictors of performance on the relatively complex tasks that are required for successful performance in real life situations. The legitimacy of these arguments is debatable given evidence that simple reaction time can be a moderately strong predictor of general intelligence (Jensen, 1982; Kennedy et al., 1985) and that intelligence, in turn, is the best overall predictor of job performance in a variety of jobs (McHenry et al., 1990; Ree & Earles, 1991). Further examination of these issues is needed to evaluate the potential value of Grammatical Reasoning as an element of field test batteries, particularly with one- and two-session research designs. Such work is important because it is possible that batteries could be constructed of cognitive and psychomotor tests which are valid indicators of abilities that are irrelevant when the objective is to predict real life performance. If complex tests are required for validity relative to the real world performance criteria and such tests do not provide psychometrically sound measures after one or two sessions of data collection, the preceding conclusion that one- and two-session research designs can be useful in this context would have to be reconsidered.

The replication of the general pattern of findings for Code Substitution and the Manikin Test using data from two additional samples was very important. Conclusions based on a single small sample must be viewed with caution, because small samples mean that chance variation in statistics can be substantial, so parameter estimates are imprecise. In addition, small chi-squares can be expected because chi-squares are the product of the fit of the model to the data and the sample size (Bollen, 1989; Hoelter, 1983; Marsh, Balla, & McDonald, 1988). The replication findings indicated that comparable results could be obtained in independent samples performing the same test. The replication findings also were consistent with evidence that parameter values for structural equations can be recovered accurately even in small samples (McArdle, 1991). While the replication of structural coefficients across two small samples for two tests was encouraging and may indicate the general robustness of the models described here, additional empirical tests are needed to further evaluate this issue. In the future, it would be desirable to be able to analyze the data in terms of covariance matrices rather than the correlation matrices analyzed in this study. The present results basically imply approximately equal proportions of true score and error variance across the samples, but did not permit direct tests of the equality of the variances themselves.

Two additional problems which lie outside the scope of this study must be addressed to accurately interpret the scores obtained. First, the present analyses utilized information about the variation around sample means on each trial and did not deal with changes in the mean value which define group learning curves. Learning curves can be expected in most research and will require additional experimental design controls, such as control groups and/or counterbalancing of exposure to the environmental conditions of interest and some appropriate control conditions (AGARD Working Group 12, 1989). Single group designs would be feasible if the learning curve for each test under neutral conditions could be established with sufficient precision to provide a realistic null hypothesis for use in significance testing, but such curves are not available at this time.

The second problem not addressed in this study is that the basis for the constant factor loadings (lambdas) observed in the present studies has not been established. Constant loadings imply that the causal factors operating on a test are constant over the series of sessions, but these causal factors are not necessarily the cognitive ability or abilities the test is designed to measure. For example, the constancy of true score variance might be the product of differences in general intelligence and reaction time (Kennedy et al., 1985) rather than differences in a specific ability such as pattern comparison. Further study to determine the components of true score variance in these cognitive tests could be useful for the purposes of identifying stable contaminants to be measured and controlled statistically when attempting to identify the effects of a stressor on a specific ability. One value of the results of the present study is that they imply that single test designs can be used for the purposes of identifying these stable components of variance, because these factors apparently operate consistently from the first trial onward. Thus, evidence from other testing traditions where single measurements are the norm, such as typical abilities assessment studies, can be brought to bear on this problem.

The overall conclusion from this study is that the cognitive and psychomotor tests considered here are likely to be valid indicators of some aspects of cognitive functioning from the first session onward. This conclusion supports recent recommendations (AGARD Working Group 12, 1989), but the need for careful research designs to treat problems arising from low measurement precision, learning curves, and carry-over effects cannot be neglected if valid results are to be obtained.

References

Ackerman, P. L. (1987). Individual differences in skill learning: An integration of psychometric and information processing perspectives. Psychological Bulletin, 102, 3-27.

Ackerman, P. L. (1988). Determinants of individual differences during skill acquisition: Cognitive abilities and information processing. Journal of Experimental Psychology: General, 117, 288-318.

AGARD Working Group 12. (1989). Human performance assessment methods. Loughton, Essex, UK: NATO Advisory Group for Aerospace Research and Development, AGARDograph No. 308.

Baddeley, A. D. (1968). A three-minute reasoning test based on grammatical transformation. Psychonomic Science, 10, 341-342.

Benson, A. J., & Gedye, J. L. (1983). Logical processes in the resolution of orientation conflict. Farnsborough, UK: Royal Air Force Institute of Aviation Medicine, Report 259.

Bentler, P. M. (1989). EQS: Structural Equations Program Manual. Los Angeles: BMDP Statistical Software, Inc.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. Psychological Bulletin, 88, 588-606.

Bentler, P. M., & Chou, C-P. (1990). Model search with TETRAD II and EQS. Sociological Methods and Research, 19, 67-79.

Bentler, P. M., & Mooijaart, A. (1989). Choice of structural model via parsimony: A rationale based on precision. Psychological Bulletin, 106, 315-317.

Bentler, P. M., & Woodward, J. A. (1980). Inequalities among lower bounds to reliability: With applications to test construction and factor analysis. Psychometrika, 45, 249-267.

Bittner, A. C., Jr., Carter, R. C., Kennedy, R. S., Harbeson, M. M., & Krause, M. (1986). Performance Evaluation Tests for Environmental Research (PETER): evaluation of 114 measures. Perceptual and Motor Skills, 63, 683-709.

Bollen, K. A. (1989). Structural equations with latent variables. NY: Wiley.

Carter, R. C., Kennedy, R. S., & Bittner, A. C., Jr. (1981). Grammatical reasoning: A stable performance yardstick. Human Factors, 23, 587-591.

Carter, R. C., & Wolstad, J. (1985). Repeated measurements of spatial ability with the Manikin

Test. Human Factors, 27, 209-220.

Cohen, J. (1968). Statistical power analysis for the behavioral sciences. NY: Academic press.

Ekstrom, R. B., French, J. W., Harmon, H. H., & Derman, D. (1976). Manual for Kit of Factor Referenced Cognitive Tests. Princeton, NJ: Educational Testing Service.

Englund, C. E., Reeves, D. L., Shingledecker, C. A., Thorne, D. R., Wilson, K. P., & Hegge, F. W. (1987). Unified Tri-service Cognitive Performance Assessment Battery (UTC-PAB). I. Design and specification of the battery. Report 87-10. San Diego, CA: Naval Health Research Center.

Hockey, G. R. J. (1986). Changes in operator efficiency. In K. Boff, L. Kaufman & J. Thomas (Eds.), Handbook of perception and performance. Vol. II (pp. 44-1-44-49). N.Y.: Wiley.

Hoelter, J. (1983). The analysis of covariance structures: Goodness-of-fit indices. Sociological Methods and Research, 11, 325-344.

James, L. R., Mulaik, S. A., & Brett, J. M. (1982). Causal analysis: Assumptions, models, and data. Beverly Hills, CA: Sage.

Jensen, A. R. (1982). The chronometry of intelligence. In R. J. Sternberg (Ed.), Advances in the psychology of human intelligence, Vol. 1 (pp. 255-310). Hillsdale, NJ: Erlbaum.

Jones, M. B. (1979). Stabilization and task definition in a performance test battery (Report NBDL-M001). Hershey, PA: Pensylvania State University.

Joreskog, K. G., & Sorbom, D. (1981). LISREL user's guide. Chicago: International Educational Services.

Joreskog, K. G., & Sorbom, D. (1990). Model search with TETRAD II and LISREL. Sociological Methods and Research, 19, 93-106.

Kennedy, R. S., Bittner, A. C., Carter, R. C., Krause, M., Harbeson, M. M., McCafferty, D. B., Pepper, R. L., & Wiker, S. F. (1981). Performance Evaluation Tests for Environmental Research (PETER): Collected papers (Report NBDL-80R008). New Orleans: Naval Biodynamics Laboratory.

Kennedy, R. S., Carter, R. C., & Bittner, A. C., Jr. (1980). A catalogue of performance evaluation tests for environmental research. Proceedings of the 24th Annual Meeting of the Human Factors Society, Los Angeles, CA, 13-17 October 1980.

Kennedy, R. S., Dunlap, W. P., Jones, M. B., Lane, N. E., & Wilkes, R. L. (1985). Portable human assessment battery: stability, reliability, factor structure, and correlation with tests of intelligence. Orlando, FL: Essex Corporation.

Klein, R., & Armitage, R. (1979). Rhythms in human performance: 1 1/2-hour oscillations in cognitive style. Science, 204, 1326-1328.

Lord, F. M., & Novick, M. R. (1968). Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. Psychological Bulletin, 103, 391-410.

McArdle, J. J. (1991). Patterns of change in structural equation models. Paper presented at the Annual Meeting of the American Psychological Association, San Francisco, 18 August.

McHenry, J. J., Hough, L. M., Toquam, J. L., Hanson, M. A., & Ashworth, S. (1990). Project A validity results: The relationship between predictor and criterion domains. Personnel Psychology, 43, 335-354.

Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Stilwell, C.D. (1989). Evaluation of goodness-of-fit indices for structural equation models. Psychological Bulletin, 105, 430-455.

Pepper, R. L., Kennedy, R. S., Bittner, A. C., Jr., & Wiker, S. F. (1980). Performance Evaluation Tests for Environmental Research (PETER): Code substitution test. Proceedings of the Seventh Psychology in the DOD Symposium, USAF Academy, Colorado Springs, CO, 16-18 April 1980.

Ree, M. J., & Earles, J. A. (1991). Predicting trainig success: Not much more than g. Personnel Psychology, 44, 321-332.

Rosenthal, R., & Rosnow, R.L. (1984). Essentials of behavioral research: methods and data analysis. NY: McGraw-Hill.

Schmitt, N., Gooding, R. Z., Noe, R. A., & Kirsch, M. (1984). Metaanalyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. Personnel Psychology, 37, 407-422.

Spirtes, P., Scheines, R., & Glymour, C. (1990a). Simulation studies of the reliability of computer-aided model specification using the TETRAD II, EQS, and LISREL programs. Sociological Methods and Research, 19, 3-66.

Spirtes, P., Scheines, R., & Glymour, C. (1990b). Reply to comments. Sociological Methods and Research, 19, 107-121.

Sternberg, S. (1966). High-speed scanning in human memory. Science, 153, 652-654.

Thorne, D. R., Genser, S. G., Sing, H. C., & Hegge, F. W. (1985). The Walter Reed Performance Assessment Battery. Neurobehavioral Toxicology and Teratology, 7, 415-418.

Tucker, L. R., & Lewis, C. (1973). The reliability coefficient for maximum likelihood factor analysis. Psychometrika, 38, 1-10.

Wherry, R. J., Sr. (1984). Contributions to correlational analysis. San Diego, CA: Academic Press.

Footnotes

[1] Model modification results obviously depend somewhat on the criterion for introducing a modification. Given that the results at any one step can capitalize on chance due to the number of parameters which are considered, a Bonferroni test could be applied to the chi-square values for modification indices (Wherry, 1984). The critical chi-square value for model modification using this approach would vary from one step to the next because the number of parameters considered would vary. The critical value would be 9.88 for the first step given 29 constrained parameters (10 true score variance estimates, 10 error variance estimates, and 9 error covariance estimates) which might be modified. If this rule had been employed instead of the change in the ATLI, the final model adopted would have been the same for 6 of 9 measures. Another possible stopping rule for model modifications would disregard the number of individual significance tests implied at any step in the modification procedure and continue modifications until every constraint that produced a chi-square change of 3.84 or greater ($p < .05$ for 1 df) had been added. This procedure was applied to the data, and it was found that the models adopted by the stricter criterion used in this study accounted for a median of 94.1% of the chi-square improvement that would have been achieved had the more lenient stopping criterion been applied. The criterion actually employed to determine the final model adopted in this study seemed to provide a reasonable balance between the risk of Type I and Type II decision errors relative to these alternative criteria. It must be emphasized, however, that the models adopted are provisional and should be revised if further research identifies replicable areas of misfit.

[2] The decision to modify models one parameter at a time represented a trade-off between possible capitalization on chance due to the consideration of a large number of statistical tests and the potential insensitivity of decisions based on the consideration of different sets of parameters. A different approach to generating alternative models to fit the data would have been to fit structural models which represented alternative conceptual models. For example, the tau-equivalent model defined by Lord and Novick (1968) could be tested for goodness-of fit to the data by freeing the equality constraints on all of the error variances in a single step. The resulting change in fit for the model would be indicated by a chi-square with 9 degrees of freedom. This chi-square would provide a diffuse significance test in Rosenthal and Rosnow's (1984) terminology because it involves estimates of more than one parameter. Diffuse tests can be misleading if the improvement in fit of the model results from changes in just a subset of the modified parameters. In this case, the overall improvement in fit may not be large enough to be significant when the critical parameters are combined with parameters that were estimated accurately in the previous model. Even if the fit does improve significantly, the test will not define which parameter(s) is(are) the source of the improvement. The alternative is to employ focused significance tests, each based on a single degrees of freedom, thereby describing the parametric location of the improvement in fit precisely (Rosenthal & Rosnow, 1984). In the present instance, the belief that the location of misfit would be confined largely to the data for one or more of the early sessions was reason to avoid diffuse tests. Analyses might have been implemented by systematically modifying the constraints for the early sessions (e.g., first the error constraint for session 1, then the error constraint for session 2, etc). However, implementing this procedure would have been equivalent to assuming that the locations of

deviations from the parallel tests model were known a priori. Among other things, this approach would have meant that it was necessary to decide whether to free constraints on the error terms or the lambda terms before examining the data. Such a priori choices would be equivalent to giving one alternative model priority over another without reference to the data. This type of procedure was rejected on the grounds that such choices were premature and that it was desirable to test the assumption that misfit would occur in the early sessions rather than in later sessions instead of just assuming this to be the case. Considering all of the possible modifications on an equal footing appeared the most reasonable approach to dealing with the problems. This approach may capitalize on chance, but these initial assessments are intended to refine hypotheses for later studies which would attempt to replicate important trends in the present findings.

[3] The model modification criterion was applied only to the focal constraints considered in the introduction. The analyses also produced estimated modification indices for an additional 36 correlated error measures. There was no a priori basis for specifying which of these correlations would be large, and modification of the relatively simple models to explain correlations between errors on temporally separated measures (e.g. cyclical changes in individual differences in functioning) did not seem appropriate until any specific findings supporting such models were replicated. In effect, those modifications which could be explained by differences between reasonably simple theoretical models were given more weight in the decision making process than findings which were of lower a priori probability. In this sense, the model construction was an exercise in Bayesian decision making with its emphasis on prior probabilities, rather than a straightforward application of fixed statistical decision criteria.

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION
Unclassified

1b. RESTRICTIVE MARKINGS
N/A

2a. SECURITY CLASSIFICATION AUTHORITY
N/A

2b. DECLASSIFICATION/DOWNGRADING SCHEDULE
N/A

3. DISTRIBUTION/AVAILABILITY OF REPORT
Approved for public release; distribution unlimited.

4. PERFORMING ORGANIZATION REPORT NUMBER(S)
Report No. 92-8

5. MONITORING ORGANIZATION REPORT NUMBER(S)

6a. NAME OF PERFORMING ORGANIZATION
Naval Health Research Center

6b. OFFICE SYMBOL *(If applicable)*
232

7a. NAME OF MONITORING ORGANIZATION
Chief, Bureau of Medicine and Surgery

6c. ADDRESS *(City, State, and ZIP Code)*
P. O. Box 85122
San Diego, CA 92186-5122

7b. ADDRESS *(City, State, and ZIP Code)*
Navy Department
Washington, DC 20372-5120

8a. NAME OF FUNDING/SPONSORING ORGANIZATION Naval Medical Research & Development Command

8b. OFFICE SYMBOL *(If applicable)*

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER

8c. ADDRESS *(City, State, and ZIP Code)*
NNMC
Bethesda, MD 20889-5044

10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO.
62233N	MM33B30	001	6104

11. TITLE *(Include Security Classification)*
(U) Field Applications of Cognitive Assessment Batteries: Initial Tests of Alternative Measurement Models

12. PERSONAL AUTHOR(S)
Vickers, Jr., Ross R., and Kusulas, Jeffrey W.

13a. TYPE OF REPORT
interim

13b. TIME COVERED
FROM ________ TO ________

14. DATE OF REPORT *(Year, Month, Day)*
91 SEP 10

15. PAGE COUNT
30

16. SUPPLEMENTARY NOTATION

17. COSATI CODES

FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*
cognitive assessment
structural modeling
military personnel
cognitive performance
psychomotor tests

19. ABSTRACT *(Continue on reverse if necessary and identify by block number)*

(U) Portable computerized cognitive assessment batteries provide a method of assessing cognitive performance during real life exposure to demanding situations. The present study evaluated the effect of relaxing guidelines regarding pre-testing to achieve valid assessments in repeated measures studies. Data reported by Kennedy et al. (1985) were reanalyzed using structural equation modeling procedures. The findings demonstrated that 7 of 9 commonly used cognitive tests could be interpreted as valid measures the first trial onward. This conclusion also may apply to Sternberg's (1966) memory test if an arguably chance finding of a change in true score variance on a single trial fails to replicate. Baddeley's (1968) Grammatical Reasoning test required a two-dimensional model to represent the data. The results for two tests (Code Substitution and Manikin) were replicated with data from other published sources. If these initial results replicate on further study, the cognitive measures examined can be used in research designs such as simple pre-post and experimental-control group research designs even when no practice sessions are feasible.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT
□ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. □ DTIC USERS

21. ABSTRACT SECURITY CLASSIFICATION
Unclassified

22a. NAME OF RESPONSIBLE INDIVIDUAL
Ross R. Vickers, Jr.

22b. TELEPHONE *(Include Area Code)*
(619) 553-8454

22c. OFFICE SYMBOL
232

83 APR edition may be used until exhausted
All other editions are obsolete

SECURITY CLASSIFICATION OF THIS PAGE

☆U.S. Government Printing Office: 1986—607-047